Contextual Anomaly Detection: looking for infrared excess in Sun-like stars

Gabriella Contardo

SMASH MSCA Fellow @ University of Nova Gorica (Slovenia)

https://arxiv.org/abs/2403.18941 (Astronomical Journal, 2024)

SMASH
machine learning for science and humanities postdoctoral program

UNIVERZA V NOVI GORICI SCIENTIA VINCES 1995

Anomaly Detection – On finding odd things

Outlier detection:

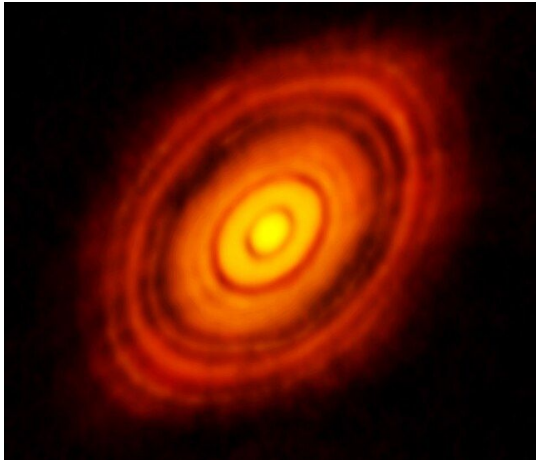
- Rare objects: points with a low probability, or in a low density region of the data space
- **“Unsupervised” learning problem**, using (some form of) density estimation
    - Obtaining reliable density estimates is non-trivial, especially in high-dimension.
    - Will retrieve all (potentially) rare objects, but not necessarily the “interesting” ones.
    - Advantages: Relevant for **unknown unknowns.**

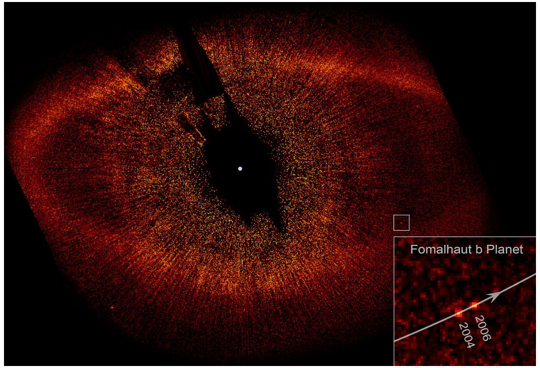
Anomaly Detection - On finding odd things

- Sometimes, we look for **“known” unknowns**:
    - Objects that are anomalous in a specific way / region of the data space : **conditional or contextual anomalies**
- Reduce the search space
- Can help our search by framing it back into a supervised problem, without technically needing supervised anomalies.

Infrared-Excess in Stars

- Infrared Excess in stars: dust (protoplanetary disks, debris disks)
- Some IR excess are more unusual: “**Extreme Debris Disks”**
  - IR fractional luminosity lower than protoplanetary disks, but >> than regular debris disks
  - A very short-lived stage in the disk evolution, or plannetary collisions? (Or dyson sphere? 👽)
  - Rare occurrence: ~0.01% (~20 candidates) from previous searches





Finding Infrared-Excess in Stars

- Usual ingredients for an IR-excess search:
  - Optical to IR observations : Often stringent quality / SNR cut in the IR.
  - A way to estimate an excess: Proper stellar model fitting (computationally expensive), template approximations. Requires correcting for reddening, models, assumptions, etc.
- Our pipeline:
  - Use mid-IR for determining the excess: much more data
  - Model MIR-emission (from optical and other features) with Machine Learning. **Anomalies according to the data**, not to a stellar model.

Anomaly Criterion Cuts

- Ensemble of predictors, combined to give an estimate of MIR-emission.
- High prediction errors: anomalous (either excess or deficit)
- We want **highly confident** incorrect predictions:

Additional criteria:

1. Low variance in prediction across models' ensemble.

$$\text{fold-MAD}_i = \text{Median}(\{|\widehat{W}_i - \widetilde{W}_{i,j}|\}_{j \in \mathrm{F}_i})$$

2. Are in *well-predicted* regions of the feature space

$$\text{K-MAD}_{i,j} = \text{Median}(\{|W_k - \widetilde{W}_{k,j}|\}_{k \in \text{NN-colour}(i,j)})$$

3. Are in a *well-populated* region of the dataset

Anomaly Criterion Cuts – Additional cuts

- We add a serie of check to prevent potential false detection
- Lead to **53 candidates** (out of 4.9M)

Criterion cut	Number of remaining candidates
Prediction Error cut (Eq. 3)	385
Mean (Prediction Error / k-MAD) cut (Eq. 4)	339
Error cut AND error/k-MAD cut	170
fold-MAD cut (Eq. 5)	127
Crowding cut at 5 arcsecond	87
FoM > 4 cut	87
Proper-Motion disagreement cut	78
Disagreement *AllWISE*/*unWISE* cut	76
Mean Distance k-NN $< .1$	66
$abs(b) > 10$	59
Removing binaries and binaries candidates (*Gaia*, Simbad)	55
Removing duplicated sources (*Gaia* DR3 flag)	53

Conclusion

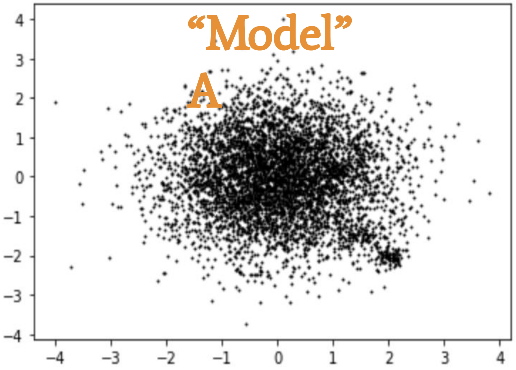
- A contextual anomaly detection pipeline
- Finding Sun-like stars with MIR-excess according to the data (and ML)
- < 100 candidates in 5M stars

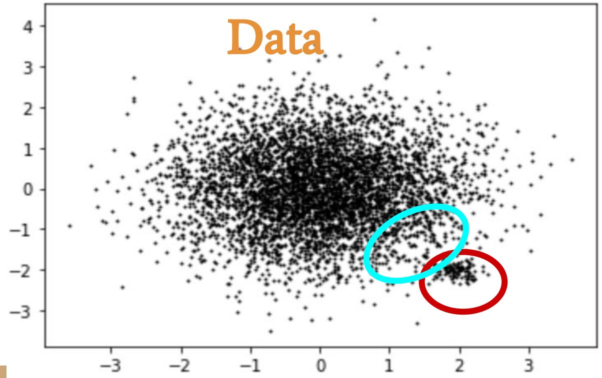
Next:

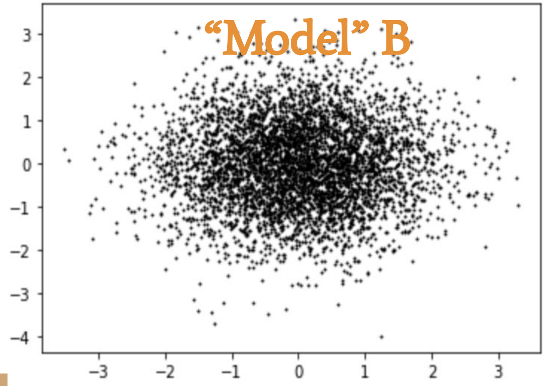
- Follow-up observations:
  - Disentangling underlying causes of excess
  - Better (?) age estimates
- Same search for other stellar types: comparing the rates and properties

Digression on the concept of Anomalies

- Some scientific communities focuses on **outlier detection** (rare objects)
- Other communities interested in finding ***divergences in distribution*** **between observations and "model"**
- Outlier only: missing interesting anomalies? But what if you don't have a model?
- Topographical features, class discovery, dimensionality reduction, ...







Thank you! Questions?